ABSTRACT

Speech nasality disorders are characterized by abnormal resonance in the nasal cavity. Hypernasal speech is of particular interest, characterized by an inability to prevent improper nasalization of vowels, and poor articulation of plosive and fricative consonants, and can lead to negative communicative and social consequences. It can be associated with a range of conditions, including cleft lip or palate, velopharyngeal dysfunction (a physical or neurological defective closure of the soft palate that regulates resonance between the oral and nasal cavity), dysarthria, or hearing impairment, and can also be an early indicator of developing neurological disorders such as ALS. Hypernasality is typically scored perceptually by a Speech Language Pathologist (SLP). Misdiagnosis could lead to inadequate treatment plans and poor treatment outcomes for a patient. Also, for some applications, particularly screening for early neurological disorders, the use of an SLP is not practical. Hence this work demonstrates a data-driven approach to objective assessment of hypernasality, through the use of Goodness of Pronunciation features. These features capture the overall precision of articulation of speaker on a phoneme-by-phoneme basis, allowing demonstrated models to achieve a Pearson correlation coefficient of 0.88 on low-nasality speakers, the population of most interest for this sort of technique. These results are comparable to milestone methods in this domain.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Nasality

Speech nasality describes the distinction between resonance in the nasal and oral cavities. Hypernasality is an inability to properly control resonance and airflow between the two cavities due to velopharyngeal dysfunction. (VPD) In an unaffected speaker, the velum, or soft palate, closes fully against the pharyngeal wall when a non-nasal sound is being produced, and opens to allow nasal resonance when needed. Velopharyngeal dysfunction refers to any inability to make this happen, and can arise from a cleft palate, hearing impairment, dysarthria, or other neurological disorder such as ALS.

Hypernasality is typically scored perceptually by a trained speech language pathologist (SLP). Due to the wide range of conditions that can give rise to hypernasal speech, it is critical that the degree of a speakers nasality be scored accurately and consistently. There are many cases where a speech language pathologist wouldnt be available but nasality scoring would be desired, such as during a check-up screening

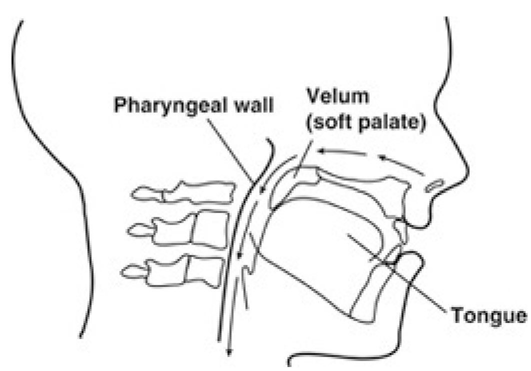

Figure 1.1: An illustration of the velopharyngeal system, responsible for controlling resonance between the oral and nasal cavities.

for early signs of neurological disorders, or in a post-op assessment following a surgery to correct a cleft palate. In situations like these automated nasality scoring would be particularly helpful. Currently, automated nasality detection systems leave a lot to be desired. This work seeks to apply a yet-unused measure to nasality prediction, the Goodness of Pronunciation score.

Some phonemes, the simplest perceptually distinct sounds in a language, are more susceptible to nasalization than others. The poor velopharyngeal closure that characterizes hypernasal speech affects the consonants that require building intraoral pressure, or expelling air out the mouth by causing leakage through the nose. This affects the stops (P, B, T, D, K, G) and the fricatives (S, Z, SH, ZH, TH, F, V). The nasal consonants (M, N, NG) are unaffected by hypernasal VPD (Meredith (2005)). Vowels undergo nasalization, which can be measured as a flattening of the space between the first and second formants, but in English nasalized vowels and non-nasalized vowels are perceived to be the same (Arai (2004)).

1.2 Goodness of Pronunciation

3.2 Goodness of Pronunciation Goodness of Pronunciation (GoP) is a score that captures a speakers articulatory precision on each individual phoneme in a spoken passage. The GoP algorithm was originally developed for use as a phoneme-by-phoneme accent quality metric for second language learners, designed to facilitate live, automated feedback on the correctness of their pronunciation against a standard phonetic language model and a transcript of the spoken passage. Specifically, it generates a time-normalized posterior log likelihood of the actual spoken phoneme given the correct phoneme from the transcript. Witt (1999)

In automated speech recognition (ASR) a frame is a unit over which the spoken phoneme is analyzed. The GoP algorithm works by first forcing an alignment between

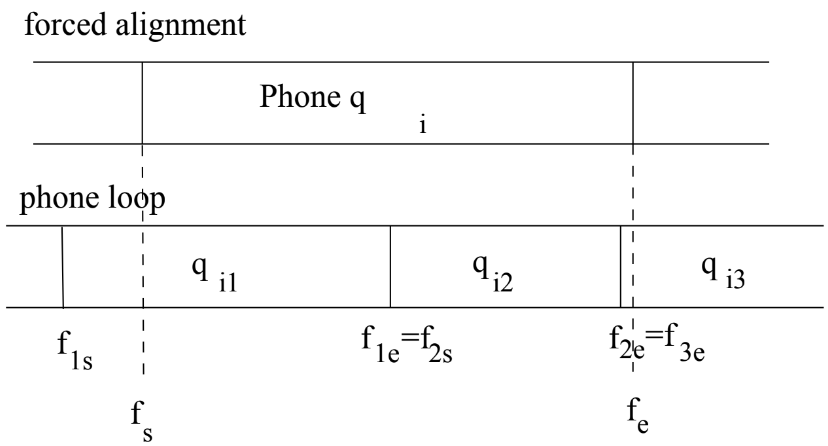

Figure 1.2: The Goodness of Pronunciation aligned frame

the transcript-generated correct phonemes and the speech passage, and then comparing the likelihood of the output acoustic segment O given the correct transcript-generated phoneme, q_i to the maximum likelihood of O given each phoneme q_j in J, the set of phonemes predicted by the ASR model in the range of the aligned frame, normalized by the number of ASR frames in O.

$$GOP_1(q_i) = |\log\left(\frac{p(O|q_i)}{\max_{j=1}^{J} p(O|q_j)}\right)|/NF(O)$$

1.3 Problem Statement

Can a simple model built on the articulatory precision measure of GoP accurately rate spoken nasality in a mild nasality range? Can a data-driven approach find the critical GoP features for solving this problem? This problem is of particular interest in the situations discussed above where accurate nasality reporting in the mild nasality range would be very useful but SLPs wouldnt be available.

The prediction is that because GoP features capture an overall success in correctly articulating a given sound, the systematic nature of shifts in nasalization can be captured by a simple model using them as input features.

Chapter 2

RELATED WORKS

2.1 Spectral Analysis of Nasality

Pruthi and Espy-Wilson demonstrate acoustic parameters for automatic vowel nasalization detection based on pole-zero pairs and measures of formants, including amplitude reduction, bandwidth increase, and spectral flattening, achieving an accuracy of 96% on the most favorable dataset to 70% on the least favorable dataset, using the acoustic parameters with an SVM classifier (Pruthi and Espy-Wilson (2007)). The limitation of this approach is that formant analysis only applies to vowels, and VPD also affects the plosive and fricative consonants. Additionally, no attempt is made to use the features for overall speaker nasalization scoring.

Kataoka et al. scores hypernasality of children with and without cleft palate on a 6-point scale using 1/3-octave spectral analysis. Analysis of the third octave bands corresponding to the first formant 1k, 1.6k and the 1/3-octave band corresponding to the second formant 2.5kHz was shown to have a high correlation to perceptual rating(r=0.84) (Kataoka *et al.* (2001)) However, this model accuracy was demonstrated on artificially manipulated spectra in order to minimize the influence of voice source characteristics on the hypernasality. In the wild this approach will probably not perform as well.

Lee et al. demonstrate the Voice low tone to high tone ratio as another measurement of nasalization. By dividing the low frequency power into high frequency power of sound power spectrum in dB, VLHR is shown to be correlated to voice nasalization, based on an expected increase after nasal decongestant treatment (Lee *et al.* (2003))

Figure 2.1: A photograph of a child wearing the nasometer

This method is only demonstrated on otherwise healthy speakers with hyponasality from allergies or upper airway infections, not on hypernasal speech. Tsai et al. demonstrate use of this measure for assessing nasality with rho=0.81 for English and rho=0.79 for Mandarin speech (Tsai *et al.* (2012)).

2.2 Nasometer

Nasometry measures nasalance, the degree of velopharyngeal opening. This is computed using the ratio of the amplitude acoustic energy at the end of the nose with the acoustic energy at the mouth (Fletcher *et al.* (1974)). The nasometer is a commercial device to assist SLPs by providing this measure, worn over the face to measure these acoustic energies (Pentax (2016)). The nasometer has two key problems – it requires a specialized and expensive device that is awkwardly worn over the face, and it still requires a trained SLP to operate it. It is useful for speech therapy but cannot be used alone for nasality scoring.

2.3 Using Goodness of Pronunciation

Kanters et al. performed an evaluation of the GoP metric in its original intended domain, computer-aided language learning. On datasets of native and non-native Dutch speakers the performance of GoP was evaluated both in its accuracy in detecting the incorrect phonemes and the success of users in improving their accent from the corresponding advice. They found an 82% accuracy for the first task. (Kanters *et al.* (2009))

Pellegrini et al. applied the GoP algorithm to the analysis of disordered French speech arising from facial palsy and found a 70% mispronunciation detection rate with 30% false positive and false rejection rate (Pellegrini *et al.* (2014)).

There has also been work in creating alternatives to the GoP algorithm for measuring phoneme-level articulatory precision. Huang et al. propose a transfer-learning based approach that utilizes a DNN transferred from existing speech recognition models to output the GoP rather than prediction, instead of separately evaluating expected phonemes and then using simple assumptions to evaluate GoP from them like Witt proposes in the original GoP specification (Witt (1999)). They achieve better results than the vanilla GoP algorithm (Huang *et al.* (2017)).

Chapter 3

PROCEDURE

3.1 Dataset

A dataset consisting of 75 speakers of varying levels of nasality and articulatory accuracy was used. Using GoP features to evaluate the nasality of speakers on this dataset is not a trivial task – this dataset contains higher nasality speakers with relatively good articulation otherwise, as well as high nasality speakers with generally poor articulation, and low nasality speakers with poorer articulation, and various combinations in between. Overall quality of articulation and nasality score are not easily correlated on this dataset.

All individuals spoke the same set of five sentences that capture the range of English phonemes. These recording were then assessed by 14 different trained SLPs, who assigned nasality scores on a scale of 1 to 7. The average nasality score for each speaker was used as a target.

The five sentences were:

- the supermarket chain shut down because of poor management
- much more money must be donated to make this department succeed
- in this famous coffee shop they serve the best doughnuts in town
- the chairman decided to pave over the shopping center garden
- the standards committee met this afternoon in an open meeting

3.2 Preprocessing

Goodness of Pronunciation software developed with the Kaldi ASR library was used to evaluate the GoP scores of each speaker in the dataset, for each sentence. For each speaker there are then five TextGrid-aligned GoP vectors, containing the GoP of each phoneme in each word in order. These GoP scores are log-likelihoods, always less than zero. A "GoP vector" for each speaker is created by averaging the GoP of each of the 41 phonemes present in the dataset across the speaker's five sentences.

3.3 Linear Models

All of the models considered in this experiment were single-layer linear regression models using PyTorch. (torch.nn.Linear) For all models considered the output target was the SLP-assessed nasality score label. For all models considered, one-out cross validation was used.

First, a Full Model which took the entire GoP vector for each speaker as an input feature was attempted. This model used the MSE between the input and output as the loss function with an added L1 weight matrix penalty term. The weight penalty was intended to ensure that only the most important phonemes be used in the analysis.

The remaining analysis involved finding the ideal subsets of the GoP vectors to use as model input. For all of these analyses the weight penalty was removed from the loss function. Two analyses were done, following data-driven and theoretically justified approaches.

For the data-driven approaches, two methods of identifying the best four phonemes were used. First, an approach based on using the largest magnitude most important coefficients. Second, a grid search of all four-phoneme GoP subsets was undertaken to find the subset that minimizes the post-training full-dataset validation MSE loss.

For the theoretically justified approaches, sets of phonemes distinguished by linguistic characteristics were selected.

Additionally for all models, two analyses were undertaken. One analysis considered the entire set of speakers, while the other only considered the mild nasality speakers with scores below 3.5, the specific region of most interest for designing the kind of model that can detect early symptoms of neurological disorders and evaluate improved speech.

For every model analysis two evaluation metrics were taken. A scatter plot showing the expected vs the predicted nasality score was taken as well as the Pearson correlation coefficient, which works well as an evaluation metric because the perfect model would achieve equal expected and predicted nasality and be very close to a linear relationship.

Chapter 4

RESULTS

4.1 Full Model

For the full model with all speakers, the Pearson correlation coefficient (PCC) was found to be $r = 0.83$.

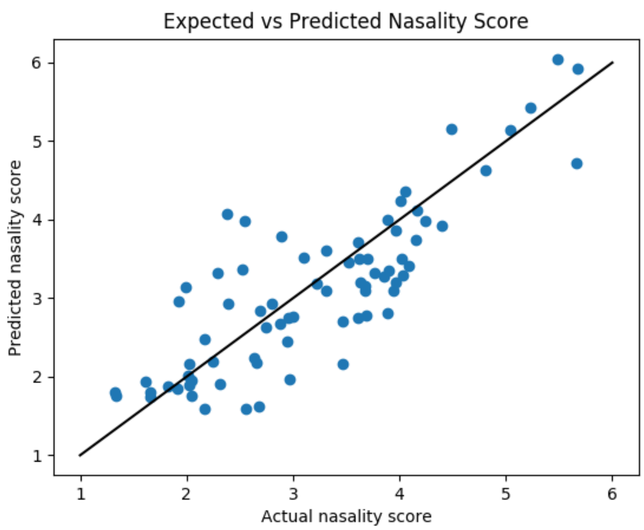

Figure 4.1: Scatter plot for the full model on all speakers.

AH	-0.18778
M	-0.22603
N	-0.21919
T	-0.47449

Table 4.1: Top four coefficients from the full model weight matrix

For the full model with mild nasality speakers, the PCC: $r = 0.88$.

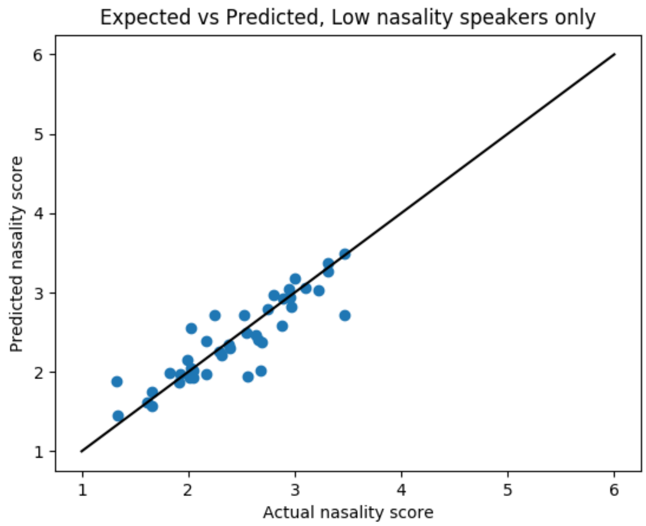

Figure 4.2: Scatter plot for the full model on mild nasality speakers only.

IH	-0.28959444
EY	0.26077273
T	-0.28105956
P	-0.25964597

Table 4.2: Top four coefficients from the mild nasality-only model weight matrix

While the mild nasality-only full model fits very well to the $Y = X$ curve, it doesn't generalize to the full speaker set well.

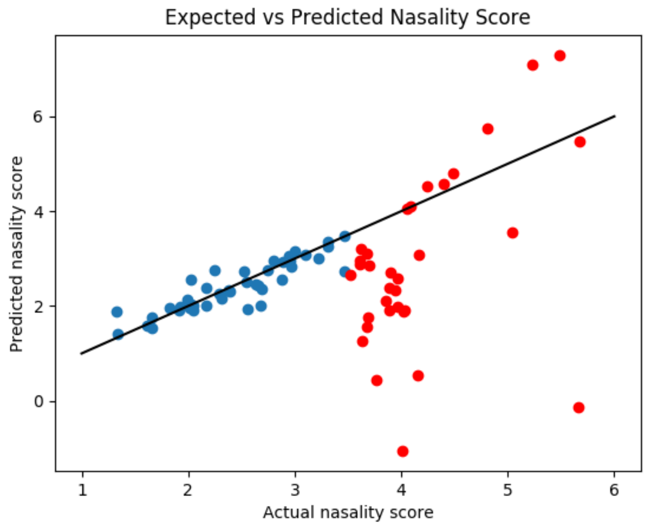

Figure 4.3: Scatter plot for the full model, trained on low nasality speakers and evaluated on all speakers.

4.2 Data-Driven Models

4.2.1 Top-4 Linear Coefficients

The top four linear coefficients from the full linear model were T, M, N, and AH.

PCC for Top-4 model with all speakers: $r = 0.72$.

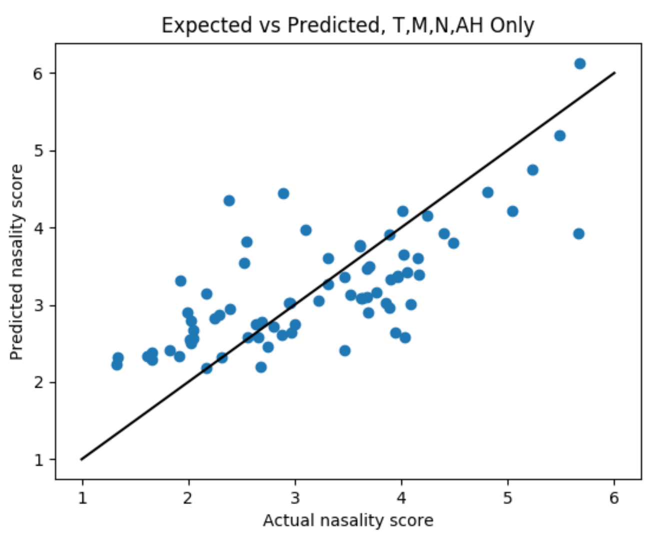

Figure 4.4: Scatter plot for the Top-4 model on all speakers.

PCC for Top-4 model with mild nasality speakers: $r = 0.46$.

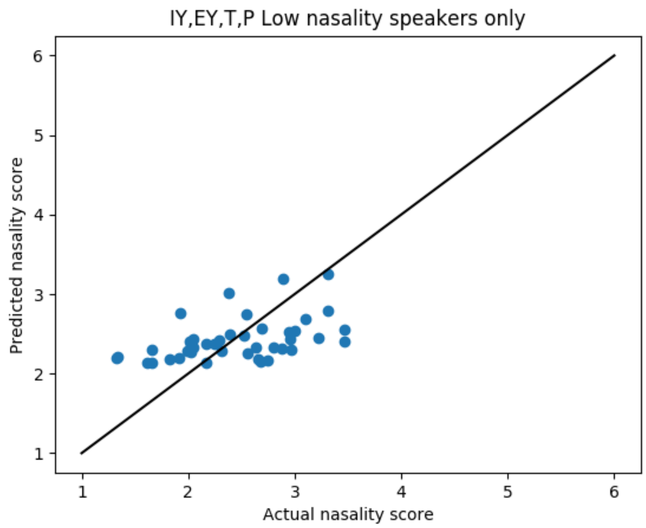

Figure 4.5: Scatter plot for the Top-4 model on mild nasality speakers only.

4.2.2 Grid Search 4

The best four phonemes for the full dataset according to the grid search were JH, OW, T, and N.

PCC for GS-4 model with all speakers: $r = 0.77$.

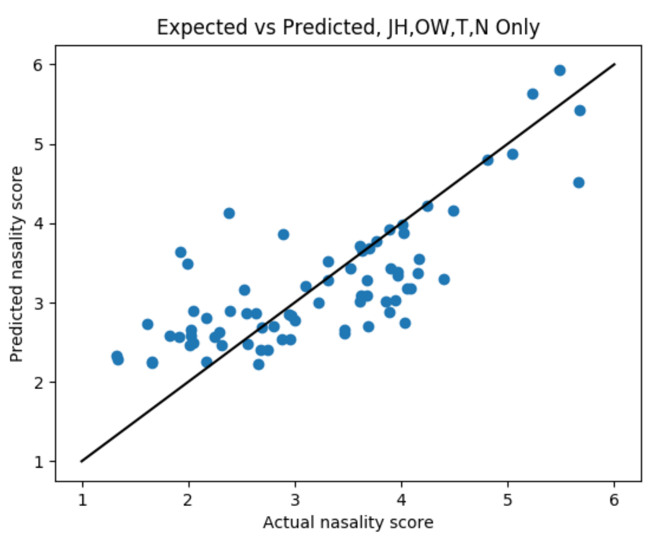

Figure 4.6: Scatter plot for the GS-4 model on all speakers.

The best four phonemes for the mild nasality speakers subset according to the grid search were R, OW, T, and M.

PCC for GS-4 model with all speakers: $r = 0.51$.

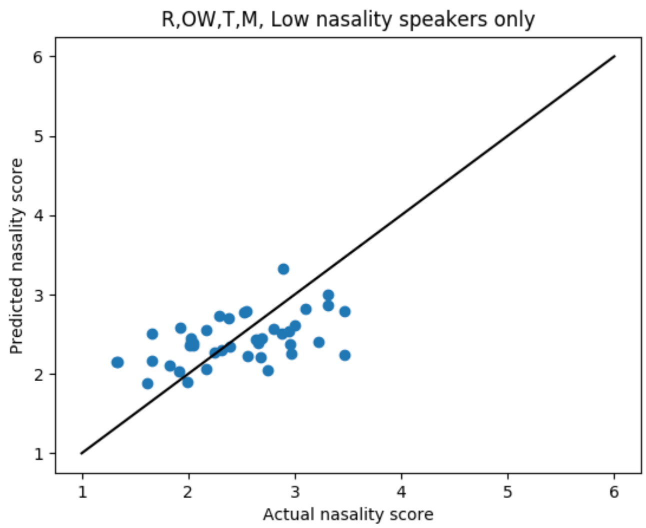

Figure 4.7: Scatter plot for the GS-4 model on the mild nasality speakers.

4.3 Theoretically Justified Models

4.3.1 Four Paired Consonants

Bilabial and denti-alveolar nasals (M and N respectively) correspond to bilabial and dental voiced plosives (B and D respectively) depending on whether the velum is closed or not. Thus the four composed the first theoretically informed subset considered.

PCC for PC-4 model with all speakers: $r = 0.64$.

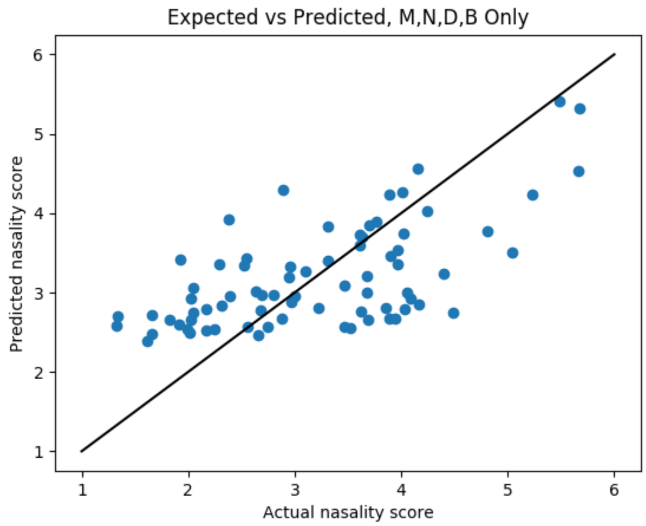

Figure 4.8: Scatter plot for the PC-4 model on all speakers.

PCC for PC-4 model with mild nasality speakers: $r = 0.48$.

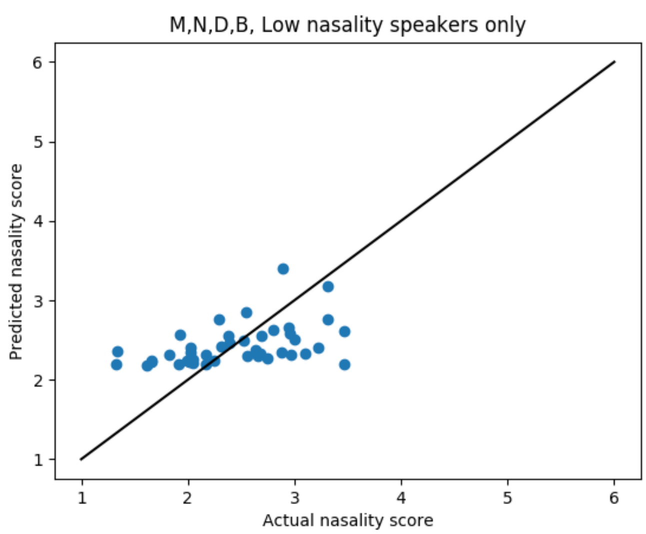

Figure 4.9: Scatter plot for the PC-4 model on mild nasality speakers.

4.3.2 Stops

Because the stops, (P, B, T, D, K, G) require a buildup of intraoral pressure, they were used as the next theoretically justified subset.

PCC for stops only model with all speakers: $r = 0.68$.

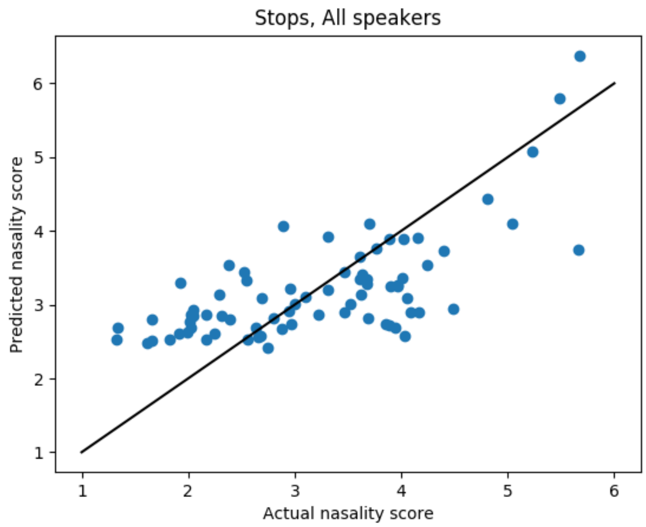

Figure 4.10: Scatter plot for the stops only model on all speakers.

PCC for stops only model with mild nasality speakers: $r = 0.48$.

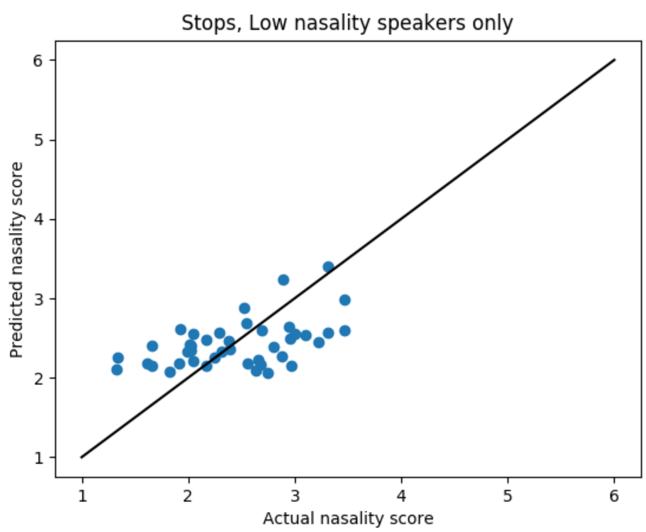

Figure 4.11: Scatter plot for the stops only model on mild nasality speakers.

4.3.3 *Fricatives*

Because fricatives, (S, Z, SH, ZH, TH, F, V) are degraded by the nasal leakage arising from VPD, they were also considered as a theoretically justified subset.

PCC for fricatives only model with all speakers: $r = 0.68$.

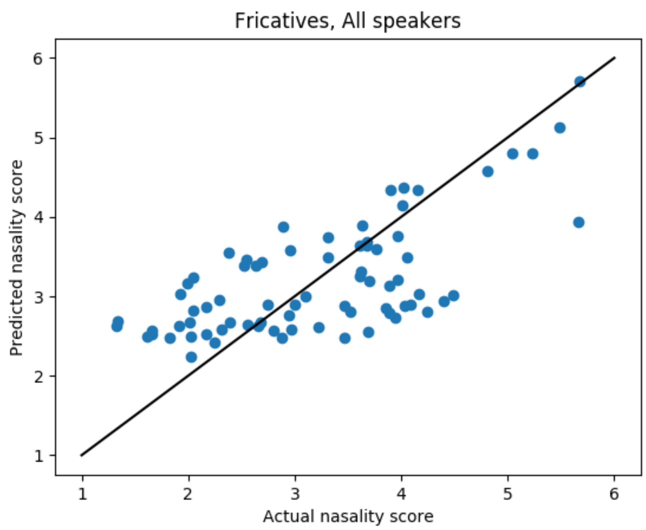

Figure 4.12: Scatter plot for the fricatives only model on all speakers.

PCC for fricatives only model with mild nasality speakers: $r = 0.42$.

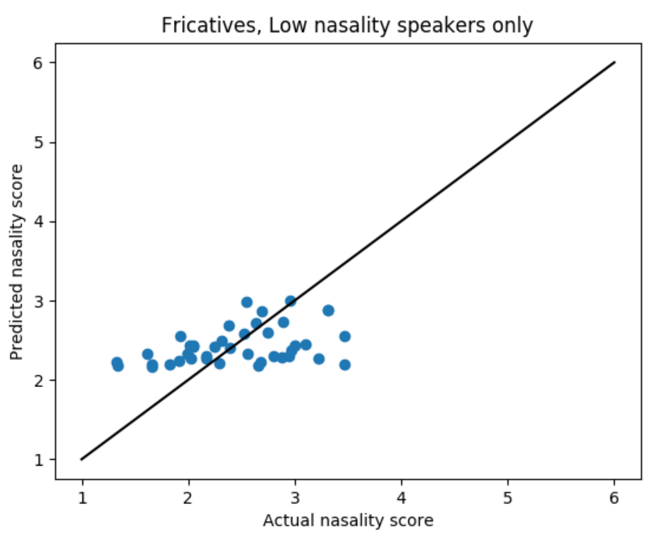

Figure 4.13: Scatter plot for the fricatives only model on mild nasality speakers.

4.3.4 Stops and Fricatives

The final theoretically-justified subset was the full set of all consonants most affected by VPI, the stops and fricatives.

PCC for fricatives only model with all speakers: $r = 0.73$.

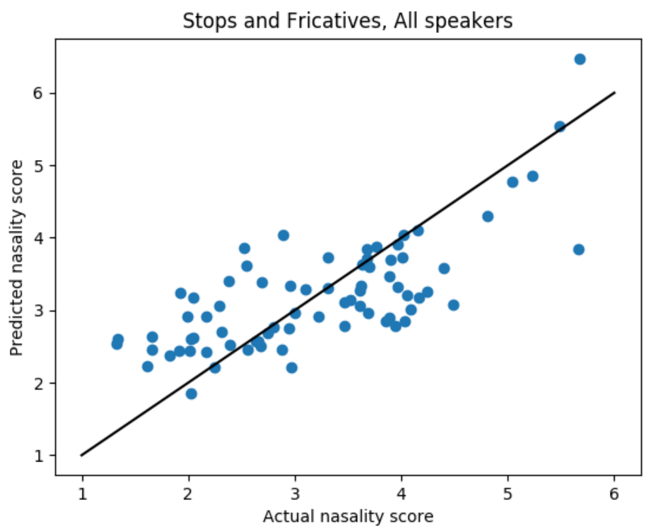

Figure 4.14: Scatter plot for the fricatives only model on all speakers.

PCC for fricatives only model with mild nasality speakers: $r = 0.63$.

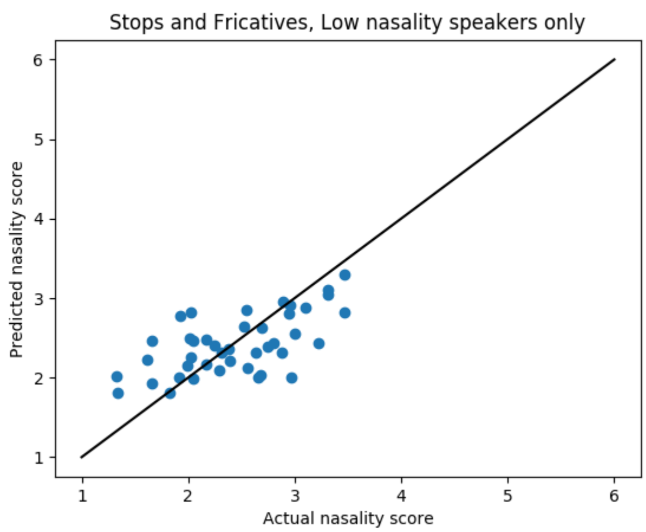

Figure 4.15: Scatter plot for the fricatives only model on mild nasality speakers.

4.3.5 Comparison of Overall Model Performances

Model	PCC all	PCC mild
Full Linear	0.83	0.88
Data-driven		
Top-4	0.72	0.49
GS-4, best	**0.77**	0.51
Theoretically justified		
PC-4	0.64	0.48
Stops	0.68	0.48
Fricatives	0.68	0.42
Stops + Fricatives	0.73	**0.63**

Table 4.3: Table of all considered models and their Pearson correlation coefficients

Bold number corresponds to best in column of non full-linear models.

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusion

These simple GoP features have been demonstrated to be effective for nasality score prediction. The best phoneme subset approaches for the full and mild nasality-only speaker sets achieve a PCC of 0.77 and 0.63 respectively, and subjectively appear to be fairly close to the perfect accuracy y=x line on the scatter plots. The full GoP feature set is very effective, particularly on the mild-nasality range subset, with a scatter plot very close to the y=x line and an r-value of 0.88.

The expectation that the articulatory precision of plosives and fricatives, the consonants most susceptible to the leakage associated with VPD, would be the most important features for nasality analysis held. In particular, the dental stop T was the highest coefficient in the full model and a part of most grid searches.

One initially surprising insight was the relative unimportance of the GoP of vowels. Only four vowels appear in the top 12 full model coefficients, and no grid-search determined subset contains more than one vowel. Although vowel sounds do become nasalized in hypernasal speech, nasalized vowels generally arent perceptually distinct phonemes from their non-nasal counterparts in the English language. Thus the English ASR model used in the GoP implementation will not distinguish them, and will assign a higher GoP score for the vowels that undergo nasalization than was expected.

For most of the evaluated models, the overall correlation coefficient of the output of the model drops when the high-nasality speakers are removed. Although the overall accuracy of the model increases, the removal of these outlying points makes the average proximity to the trendline through them higher, with the notable exception of the full all-phones linear model. In this case the scatter plot moves very close to approximating the y=x line and the PCC jumps from 0.83 to 0.88 when the high nasality speakers are removed. Because there are only 46 speakers in the mild nasality range, and 42 parameters in the full linear model, it is possible that something closer to memorization than learning the underlying function is taking place in this case.

Overfitting probably is not taking place, as on all models there is no characteristic rise in validation error as training error decreases.

However, for all models the performance of the estimation on the mild nasality range improves with the removal of high-nasality speakers, even as the PCC drops. This improvement could be related to the way that the MSE error metric is disproportionately affected by outliers in the dataset, and a fidelity in identifying the way that the articulatory precision of mildly nasal speakers relates to their nasality rating might be compromised by also accounting for high-nasality speakers. Put another way, there may not be a simple linear articulatory-precision-to-nasality relationship across the entire nasality score range, and the way that overall articulatory precision affects the nasality score needs further analysis by the provided, separately SLP-rated articulatory precision measure that is also available for each speaker. The idea that there are deeper differences in the way that articulatory precision is related to nasality is supported by the poor generalization of models trained on the mild nasality speaker set to the high nasality speakers.

Another takeaway from this analysis is that fundamentally, the utility of limited feature sets is limited, or at least, the naive approach to selecting these limited subsets hurts generalization. This is reflected in the generally poorer performance of smaller feature sets against the larger ones. This is probably related to the general effect of being able to get a better fit from having a larger input feature vector. Future analysis should involve searching through larger GoP subsets to confirm that, for example the stops+fricatives set is near ideal amongst its peers of larger subsets.

Finally, it is important to note that these models are using what is ultimately a subjective rating as a target. There is a level of baked-in difficulty in finding an objective measure to match a perceptual, subjective target. Although the averaged target nasality scores do reflect a consensus between multiple SLPs, the utility of relative nasality ratings between smaller and smaller ranges becomes less and less meaningful.

5.2 Future Work

Developing a statistical model that estimates the articulated phones without needing a matched transcript, and comparing the high-level statistics of the detected phonemes between speakers might yield an overall relationship between these statistics and nasality. However, this isn't a huge advantage as a patient can just be asked to read aloud a given sentence during the test.

Additionally, the GoP algorithm can be modified to, rather than estimating the overall probability that a phoneme was correctly identified given an audio sample, that a phoneme was mistaken for the corresponding cognate (dental and alveolar plosives to /n/, bilabial and labiodental plosives to /m/).

Evaluating performance of the model using a DNN-based GoP metric such as is detailed in Huang et al. rather than the GMM-based model for phone-level articulatory precision estimation might also be fruitful.

Finally, combining this approach, which has demonstrated success in finding the correlation between nasality score and articulation of certain consonants in particular could be combined with a spectral analysis method such as is detailed in Kataoka et al. or Lee et al.

REFERENCES

Arai, T., "Formant shift in nasalization of vowels", The Journal of the Acoustical Society of America **115**, 5, 2541–2541, URL `https://doi.org/10.1121/1.4783589` (2004).

Fletcher, S. G., I. Sooudi and S. D. Frost, "Quantitative and graphic analysis of prosthetic treatment for "nasalance" in speech", Journal of Prosthetic Dentistry **32**, 3, 284–291, URL `http://dx.doi.org/10.1016/0022-3913(74)90032-8` (1974).

Huang, H., H. Xu, Y. Hu and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection", The Journal of the Acoustical Society of America **142**, 5, 3165–3177, URL `https://doi.org/10.1121/1.5011159` (2017).

Kanters, R., C. Cucchiarini and H. Strik, "The goodness of pronunciation algorithm : a detailed performance study", in "In SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education", pp. 2–5 (2009).

Kataoka, R., D. W. Warren, D. J. Zajac, R. Mayo and R. W. Lutz, "The relationship between spectral characteristics and perceived hypernasality in children", The Journal of the Acoustical Society of America **109**, 5, 2181–2189, URL `https://doi.org/10.1121/1.1360717` (2001).

Lee, G., C. Yang and T. Kuo, "Voice low tone to high tone ratio–a new index for nasal airway assessment", The Chinese journal of psyiology **46**, 3, 123–7, URL `https://www.ncbi.nlm.nih.gov/pubmed/14672281` (2003).

Meredith, A. S., "Speech sounds affected by vpi amy skinder meredith", URL `https://www.speechpathology.com/ask-the-experts/speech-sounds-affected-by-vpi-1103` (2005).

Pellegrini, T., L. Fontan, J. Mauclair, J. Farinas and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech", in "The 15th Annual Conference of the International Speech Communication Association - INTERSPEECH 2014", pp. pp. 1463–1467 (International Speech Communication Association (ISCA), Singapore, SG, 2014), URL `http://oatao.univ-toulouse.fr/13139/`.

Pentax, "Nasometer ii: Model 6450", URL `https://www.pentaxmedical.com/pentax/en/99/1/Nasometer-II-Model-6450` (2016).

Pruthi, T. and C. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization.", (2007).

Tsai, Y., C. Wang and G. Lee, "Voice low tone to high tone ratio, nasalance, and nasality ratings in connected speech of native mandarin speakers: a pilot study.", The Cleft Palate-Craniofacial Journal **49**, 4, 437–46, URL `https://www.ncbi.nlm.nih.gov/pubmed/21740178` (2012).

Witt, S. M., *Use of speech recognition in computer-assisted language learning*, Ph.D. thesis, University of Cambridge (1999).

APPENDIX A

PROCESSING CODE

APPENDIX B

MODEL CODE